DOCUMENT RESUME

ED 107 720 TM 004 577

AUTHOR Sauls, Judith M.; Larson, Robert C.
TITLE Exploring National Assessment Data Using Singular Value Decomposition.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
PUB DATE [Apr 75]
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; Age Differences; Community Characteristics; Demography; *Educational Assessment; Geographic Regions; *Matrices; Measurement Techniques; *National Surveys; Orthogonal Rotation; Parent Education; *Performance Factors; Probability; Race; Sex Differences; Testing
IDENTIFIERS *National Assessment of Educational Progress; Singular Value Decomposition

ABSTRACT

National data was obtained from 9-year-old, 13-year-old, 17-year-old, and 26 through 35-year-old populations in order to determine academic achievement in nine subject areas. For each age population, group data was calculated and reported by region, sex, color, parents' educational level, and size and type of community. The application of singular value decomposition of nonsquare matrices to this data is described and its relationship to principal components analysis and its data reduction value is explained. Exploratory analyses are being conducted to determine if the same bases occur across age levels, across time from one assessment to its reassessment, and across subject areas. Emphasis will remain on trying to relate the characteristics of exercises to major differences in performance through the use of orthogonal components. (Author/BJG)

ED107720

EXPLORING NATIONAL ASSESSMENT DATA USING SINGULAR VALUE DECOMPOSITION

Judith M. Sauls and Robert C. Larson

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

National Assessment of Educational Progress
The Education Commission of the States

Paper Presented at 1975 Convention of
American Educational Research Association

Session 21.20/Multivariate Analysis
(Critique, Division D)

TM 004 527

EXPLORING NATIONAL ASSESSMENT DATA USING SINGULAR VALUE DECOMPOSITION

Judith M. Sauls and Robert C. Larson

National Assessment of Educational Progress
Education Commission of the States

Introduction to National Assessment

The National Assessment of Educational Progress (NAEP), a project of the Education Commission of the States, was established to collect reliable information on the achievement of educational outcomes by assessing the skills, knowledges, and attitudes of America's young people. Now in its sixth year of field assessment, National Assessment has obtained data in nine different subject areas: Science, Citizenship, Writing, Reading, Literature, Social Studies, Music, Mathematics, and Career and Occupational Development. Both Science and Writing have been assessed twice. All exercises administered by NAEP are based on important educational goals as determined by panels of educators, scholars, and laymen. Each subject area is assessed periodically in order to determine growth or decline in educational attainment.

National data is obtained for four age populations -- 9-year-olds, 13-year-olds, 17-year-olds, and young adults aged 26-35; these four age levels represent the end of primary, intermediate, secondary, and post-secondary education. Within each age, group data is calculated and reported by region, sex, color, parents' educational level, and size and type of community.

A national probability sample is used to identify respondents resulting in about 2100-2200 respondents per exercise. Each respondent takes only a portion of the total number of exercises administered at an age level. Scores for individuals are not obtained. The exercise, rather than the individual, is the basic unit of interest. For each exercise, NAEP estimates the proportion of people at a given age who can complete the exercise successfully. Using the national probability sample, it is possible to estimate national percentages correct on each exercise within certain limits of accuracy. Estimates of group performance are also obtained for each exercise. Typically, group data is reported relative to national by finding the difference between the group and national percentages correct. This difference is called the "group effect" and indicates the relative performance of the group on a particular exercise.

For example, the following exercise was given to 13-year-olds during the 1969-70 assessment of Science:

Which of the following is true of hot water as compared with cold water?

- ○ It is denser.
- ○ It is easier to see through.
- ● Its molecules are moving faster.
- ○ It has more free oxygen dissolved in it.
- ○ It has more free hydrogen dissolved in it.
- ○ I don't know.

The estimated national percentage correct for 13-year-olds was 61%. Group data were also calculated for four regions (Northeast, Southeast, Central, and West), for males and females, for Blacks and Whites, for four levels of parents' education (no high school, some high school, graduated high school, and post high school), and for size and type of community (extreme rural, inner city, affluent suburb, rest of big city, urban fringe, medium city, and small places).

For this exercise, then, the following national and group data were obtained:

	Region				Sex		Parents' Education			
NAT	NE	SE	C	W	M	F	NHS	SHS	GHS	PHS
61	2.4	-10.6	5.8	-0.7	0.6	-0.4	-16.4	-9.2	-0.9	5.9

			STOC				Color	
ER	IC	AS	RBC	UF	MC	SP	W	B
-15.1	-19.8	11.4	-2.7	-0.4	1.0	4.4	4.5	-24.4

The Data Reduction Technique of Singular Value Decomposition

Exercise level data are obtained on about 100-150 exercises per age level per subject area, resulting in a huge data base. NAEP has currently been exploring a data reduction technique, singular value decomposition (SVD), utilizing a computer algorithm for the decomposition of non-square matrices designed

by Golub and Reinsch (1970). Our purpose is to determine if any meaningful, underlying orthogonal dimensions can be found to describe the relations among exercises and among groups.

Because of NAEP's unusual data base in which different national samples of individuals respond to only a subset of the exercises, some of the usual descriptive and correlational procedures are not applicable. For example, factor analysis techniques, which attempt to explain observed relations among numerous variables in terms of simpler relations are not directly applicable to NAEP's basic percentages of correct group responses. The basic unit of interest is not the individual but the exercise and how certain groups of individuals perform on that exercise. The variables of interest to NAEP are not repeated measures on the same unit but are different classifications of the same set of respondents.

One alternative data reduction technique is the general method of singular value decomposition which obtains both left- and right-hand orthonormal characteristic vectors of a non-square matrix. NAEP is using the efficient decomposition procedure proposed by Golub and Reinsch to factor exercise by group arrays, of dimension approximately 100 X 20, whose matrix elements are the group effects associated with each exercise. For a given age level and subject area, this data array contains the basic information for an age level for any one assessment. By using SVD, we gain an economy of description--for both exercises and for groups of respondents at an age level--by obtaining

orthonormal bases for the spaces spanned by exercise vectors (rows) and by the group vectors (columns). Thus, we obtain at once information concerning the simple relations among exercises and among groups.

The purpose of this paper is to describe singular value decomposition of non-square matrices, to show its relationship to principal components analysis, and to illustrate its data reduction value for exploring National Assessment data.

The Method of Singular Value Decomposition

A statement of the basic theorem used in singular value decomposition can be found in most texts on linear algebra. Horst (1963, 1965) also gives an extensive discussion on the interpretation of this procedure, which he calls finding the basic structure of a matrix. The theorem states that a real $m \times n$ ($m \geq n$) matrix X can always be expressed as the product of three matrices:

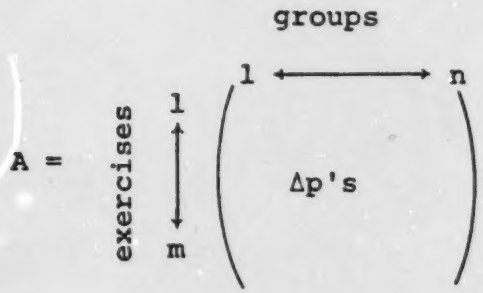
$$X = U\Sigma V'$$

where U is an $m \times n$ orthonormal matrix such that $U'U = I_n$, V is an $n \times n$ orthonormal matrix such that $V'V = VV' = I_n$, and Σ is a diagonal matrix of dimension $n \times n$. The diagonal matrix Σ contains non-negative elements, s_j for $j = 1$ to n, called the singular values, which are arranged in descending order of magnitude from upper left to lower right. The number of non-zero singular values is equal to the rank $r \leq n$ of the matrix.

The first r columns of U form an orthogonal basis for the column vectors of X and the first r columns V (rows of V′) form an orthogonal basis for the row vectors of X.

For a square, symmetric matrix, U = V and the orthonormal basis of columns and rows are same. For the well-known case of X = R, a correlation matrix, the columns of U (multiplied by the corresponding singular value) are the principal components of R and the decomposition is referred to as the principal components analysis of R. For any square symmetric matrix, the elements of the diagonal matrix Σ are equal to the square root of the characteristic roots; that is, $s_j = \sqrt{\lambda_j}$ or $s_j^2 = \lambda_j$.

When SVD is used with National Assessment data, we begin with an exercise by group data array A. Each element of A, Δp_{ij}, is the relative performance ot group j on exercise i. We can represent A as:



Next, each column of the data array A is centered about the column mean and standardized to unit variance. This standardizing procedure has been introduced to equalize the variability of the group effects across exercises. It can be shown that the variability of group effects depends on the population sizes of the

groups. Small groups show much more variability across exercises than do the larger groups. Standardizing by columns is one way to eliminate this variability and weight each group equally in the decomposition.

Conceptually, we now have a new matrix X which can be written as

$$X = U\Sigma V'$$

That is, X can be represented as:

$$X = \left[\begin{pmatrix} \\ \underset{\sim}{x}_1 \\ \\ \end{pmatrix} \begin{pmatrix} \\ \underset{\sim}{x}_2 \\ \\ \end{pmatrix} \quad \cdots \quad \begin{pmatrix} \\ \underset{\sim}{x}_n \\ \\ \end{pmatrix} \right]$$

where each column vector $\underset{\sim}{x}_j$ is m x 1. The columns of U form an orthogonal basis for the columns of X. We can represent U as:

$$U = \left[\begin{pmatrix} \\ \underset{\sim}{u}_1 \\ \\ \end{pmatrix} \begin{pmatrix} \\ \underset{\sim}{u}_2 \\ \\ \end{pmatrix} \cdots \begin{pmatrix} \\ \underset{\sim}{u}_r \\ \\ \end{pmatrix} \begin{pmatrix} \\ \underset{\sim}{u}_{r+1} \\ \\ \end{pmatrix} \cdots \begin{pmatrix} \\ \underset{\sim}{u}_n \\ \\ \end{pmatrix} \right]$$

where each column vector $\underset{\sim}{u}_j$ is m x 1. The first r columns of U form an orthonormal basis for the column vectors of X.

The total column variance of X is the sum of the n column variances, or n. The total column variance can also be expressed as:

$$\text{trace } \frac{X'X}{m} = \text{trace } \frac{V\Sigma^2V'}{m}$$

$$= \frac{1}{m} \text{trace } \Sigma^2VV'$$

$$= \frac{1}{m} \text{trace } \Sigma^2$$

$$= \frac{1}{m} \sum_{j=1}^{r} s_j^2 .$$

If we multiply each vector $\underset{\sim}{u}_j$ by its corresponding singular value s_j, we obtain $s_j\underset{\sim}{u}_j$, a vector of elements s_ju_{ij} which correspond to the usual factor loadings of exercise i on principal factor (or component) $s_j\underset{\sim}{u}_j$. The variance of each vector $s_j\underset{\sim}{u}_j$ is

$$\frac{1}{m} s_j^2 \underset{\sim}{u}_j' \underset{\sim}{u}_j = \frac{1}{m} s_j^2 .$$

The total variance of the set of $s_j\underset{\sim}{u}_j$ for j = 1 to r is $\frac{1}{m} \sum_{j=1}^{r} s_j^2$, which is equal to the total column variance of X.

Thus, the proportion of total column variance of X accounted for by the orthonormal vector $s_j\underset{\sim}{u}_j$ of X is s_j^2/mn.

One might note that $\frac{X'X}{m}$ is equal to the correlation matrix R among groups and can be factored as follows:

$$R = \frac{1}{m} (V\Sigma U')(U\Sigma V')$$

$$= \frac{1}{m} V\Sigma^2 V'.$$

The columns of V are the characteristic vectors of R and the columns of $\frac{1}{m} \Sigma V$ are the principal components of R. The characteristic roots, λ_j, of R are equal to s_j^2/m. That is, if we had begun with the correlation matrix and found $R = P\Delta P'$ $(P'P = I)$, where P contains the characteristic vectors of R and Δ is a diagonal matrix containing the characteristic roots, then the relationship between the characteristic roots λ_j of R and the singular values s_j of X is:

$$\lambda_j = s_j^2/m.$$

In both cases, the amount of variance attributable to component j of X and the corresponding component j of R is the same; both are equal to s_j^2/m. The proportion of total column variance is also the same for each component j of either X or R and is equal to s_j^2/mn.

In addition to the factor loadings of $s_j u_j$ for the columns of X, one can obtain the factor loadings for the rows of X from

the V′ matrix. V′ can be represented as:

$$V' = \begin{bmatrix} (\quad \underset{\sim}{v}_1' \quad) \\ (\quad \underset{\sim}{v}_2' \quad) \\ \\ (\quad \underset{\sim}{v}_r' \quad) \\ (\quad \underset{\sim}{v}_{r+1}' \quad) \\ \\ (\quad \underset{\sim}{v}_m' \quad) \end{bmatrix}$$

If we multiply each vector $\underset{\sim}{v}_k$ by its corresponding singular value s_k, we obtain $s_k\underset{\sim}{v}_k$ whose elements $s_k\underset{\sim}{v}_{\ell k}$ correspond to the usual factor loadings of group ℓ on component k. The amount of variance attributable to component k of ΣV is equal to s_k^2/n. The proportion of total row variance due to component k is equal to $(\frac{1}{n} s_k^2 /$ total row variance of X). Note that the total row variance is not equal to n since the rows were not centered or standardized.

Application of SVD Using National Assessment Data

The application of SVD using National Assessment data has centered on the interpretation of the column vectors of U, the orthogonal basis for the column vectors of X. Basically, we want to interpret the orthogonal components by a) correlating group performance vectors with the orthogonal components and b) considering the relative weighting of exercises on each

component and associating by inspection exercise characteristics with the orthogonal dimensions and group performance.

To illustrate how SVD is used at National Assessment, consider the data obtained for 17-year-olds in the first assessment of Science (1969-70).* One hundred twenty-four exercises were administered to several national samples of 17-year-olds. For each exercise, data were calculated and reported for groups defined by region, sex, color, parents' education, and size and type of community. Each column of the 124 x 19 data array was centered and standardized, resulting in a new matrix X which was then factored using SVD. The singular values of Σ and the percentage of total column variance of X accounted for by each component are presented in Table 1. The first three components of U account for 54% of the total column variance of X.

One way of interpreting these components is to determine how group performance correlates with the components of U. Table 2 presents the results of the correlations of columns of X with the first three columns of U. From these correlations, it appears that the first component is related to relative group standing. Large positive correlations occur for groups that typically perform above the national level of performance--the

---

* The original data matrix of group effects, the standardized matrix X, and the three matrices U, Σ, and V found from SVD can be obtained from the authors. The size of these arrays prohibits their publication in this paper.

affluent suburb, whites, and the post high school parents' education group. Large negative correlations occur for groups that typically perform below the national level--the Southeast, the extreme rural, the inner city, Blacks, and the two lowest categories of parents' educational level (no high school and some high school). Correlations close to zero occur for groups that typically perform close to the national level--Central, rest of big city, and small places.

The second component appears to be related to male-female differences. That is, after accounting for differences in relative group performance, the next largest orthogonal component separates male and female performance. For Science, males tend to typically perform better than females so that a male-female component seems logical.

The third component shows only two large correlations--a positive one for Northeast and a negative one for Central. Thus, this component appears to measure some sort of regional effect which is orthogonal to relative group performance and male-female differences.

After examining correlations of groups with components of U, we can look at the correlations of linear combinations of groups, such as male-female, to further clarify the interpretations of the components. This was done by taking differences of all pairwise columns of X and correlating these resulting vectors with columns of U. The highest correlations with component one of U were found for differences between high

performance and low performance groups. The highest correlation (-.96) was found for Black-post high school, two of the most extreme groups in terms of relative performance. High correlations were also found for Southeast-post high school (-.92), White-Black (.91), extreme rural-post high school (-.90), both none and some high school-post high school (both at -.89), and Southeast-White (-.89). These correlations tend to confirm the interpretation of the first component as relative group performance.

For component two, the largest correlation (.88) found was with the male-female vector. Other high correlations were found for male-rest of big city (.86), and female-small places (-.86). The highest correlation (.89) with component three was Northeast-Central.

Since there does seem to be linear combinations of groups that correlate highly with the components of U, another helpful analysis might be to find the linear combination of group vectors that provides the maximum possible correlation with each orthogonal component. This can be accomplished through a simple application of multiple regression analysis where the columns of X are taken as the independent variables and the orthogonal columns of U are taken in turn as the dependent variable.

Since we know that levels of parents' education are surrogates for or correlated with levels of income and type of community (STOC), it would be helpful if one could find orthogonal components that separate parents' education from income

level. Again, one can take in turn the two sets of vectors for the variables PEd and STOC as the independent sets and find the linear combination which correlates maximally with each orthogonal component.

The next step is to try to interpret the orthogonal components in terms of the exercise characteristics. For each component we can obtain an ordering, from high to low, of the exercises. Then by simple inspection it is possible to associate exercise characteristics with this ordered set of exercises to determine if there is a strong association between certain exercise characteristics and the orthogonal components. If a strong relationship exists, then one would expect exercises with common characteristics to collect at each end of the ordered orthogonal vector. In Science, for example, each exercise was characterized by type of Science (physical, biological, or other), by objective,* and by whether the correct answer might be commonly learned from a book or from another source (book/non-book category). By examining each of these ordered vectors, we found that none of these three classifications appeared to be strongly associated with any of the orthogonal components at age 17. To the extent that we have already shown that some of

---

* The objectives used in the first assessment of Science were: I. Know the fundamental facts and principles of Science, II. Possess the abilities and skills needed to engage in the processes of science, III. Understand the investigative nature of science, and IV. Have attitudes about and appreciation of scientists, science, and the consequences of science that stem from adequate understandings.

these orthogonal components are strongly associated with certain major group differences in performance, we might conclude that these differences occur on exercises regardless of science type, objective, or source of learning.

The previously mentioned characteristics are all related to content and learning. As developers and administrators of large numbers of exercises each year, NAEP is also concerned with the relation between "non-content" characteristics of exercises and differences between group performances. For example, position of correct response to a multiple choice exercise, reading level of the exercise, format (multiple choice, open-ended, or multiple choice with "I don't know" foil), position in package, time allowed to respond to the item, and so on are non-content characteristics that one usually hopes are unrelated to differences in group performance. Thus, we can use these techniques to gain some large scale item-analysis information.

Conclusion

National Assessment collects and reports much data for each assessment area. The method of singular value decomposition is being used to determine the underlying dimensions of that data base. Exploratory analyses are being conducted to determine if the same bases occur across age levels, across time from one assessment to its reassessment, and across subject areas. Emphasis will remain on trying to relate the characteristics of

exercises to major differences in performance through the use of orthogonal components.

Further development of these initial analysis procedures is also under way and can be sketched briefly. The method of association of exercise characteristics with orthogonal components can be made objective through the use of elementary correlation procedures. One might construct a matrix of exercise characteristics and obtain its orthogonal components. Then correlate these components with the exercise characteristic vectors using procedures described for correlating performance vectors. One might then correlate the exercise characteristic vectors to the performance vectors through multiple regression or canonical correlation procedures.

Another area that needs some attention is the method of standardizing the columns of the original performance matrix X. Recall we had standardized each column to unit variance, making the total variance equal to n, the number of columns. Since each of the five variables (region, sex, etc.) is really a reclassification of the same population of 17-year-olds, one might argue that the variables, rather than the groups within each variable, should be given equal weight. For example, standardizing each column to unit variance gives STOC, which has seven groups, 7/n of the total column variance, while sex with only two groups has a proportion of 2/n. Thus, the STOC variable carries three and a half times the weight of the sex variable. A simple solution to this problem is to weight each group within a variable by

multiplying by $1/\sqrt{g}$ where g is the number of groups within a variable. Then the sum of the column or group variances for each variable would be one and the total column variance for X would be equal to the number of variables.

These are just a few of the possibilities for further analysis. Clearly, the basis of these analyses is the flexibility and adaptability of SVD to data sets of different types.

Table 1

Results of SVD for 17-Year-Olds,
Science (1969-70)

Component	Singular Value	Cum. % of Total Column Variance
1	27.53	32.18
2	17.56	45.27
3	14.44	54.11
4	13.42	61.75
5	12.95	68.97
6	11.73	74.72
7	10.91	79.77
8	9.96	83.98
9	9.60	87.89
10	8.29	90.80
11	8.07	93.56
12	7.20	95.77
13	6.52	97.57
14	6.34	99.27
15	3.19	99.71
16	1.90	99.86
17	1.49	99.95
18	0.85	99.98
19	0.52	100.00

Table 2

Correlations of Columns of X with
The First Three Columns of U

Groups (Columns of X)		Components of U (Columns of U) U1	U2	U3
Region	NE	.32	-.08	.72
	SE	-.76	.27	.04
	C	.07	-.27	-.73
	W	.48	.01	-.14
Sex	M	.31	.88	.08
	F	-.30	-.87	-.17
STOC	ER	-.55	.36	-.16
	IC	-.74	-.18	.05
	AS	.62	-.05	.23
	RBC	.06	-.49	.06
	UF	.34	-.35	.38
	MC	.34	.09	-.29
	SP	-.24	.47	-.26
Color	W	.83	.07	-.28
	B	-.87	-.10	.12
Parents Ed.	NHS	-.73	-.06	.06
	SHS	-.82	.11	.08
	GHS	.17	-.16	-.32
	PHS	.91	-.04	.05

REFERENCES

Golub, G.H. and Reinsch, C. Singular value decomposition and least squares solutions. American Mathematician, 1970, 14, 403-420.

Horst, P. Matrix algebra for social scientists. New York: Holt, Rinehart and Winston, 1963.

Horst, P. Factor analysis of data matrices. New York: Holt, Rinehart and Winston, 1965.